ESD-TR-67-75

MTR-305

A SEISMIC CLASSIFICATION MODEL

SEPTEMBER 1967

J. W. Clark

Prepared for
DIRECTORATE OF PLANNING AND TECHNOLOGY
DEVELOPMENT ENGINEERING DIVISION
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

Sponsored by
Advanced Research Projects Agency
Project VELA UNIFORM
ARPA Order No. 800

Project 6050
Prepared by
THE MITRE CORPORATION
Bedford, Massachusetts
Contract AF19(628)-5165

This document has been approved for public release and sale; its distribution is unlimited.

The research reported in this document was sponsored by the Advanced Research Projects Agency, Project VELA UNIFORM, ARPA Order No. 800, under Air Force Contract No. AF-19(628)-5165 with the Electronic Systems Division, AFSC.

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Do not return this copy. Retain or destroy.

ESD-TR-67-75

MTR-305

A SEISMIC CLASSIFICATION MODEL

SEPTEMBER 1967

J. W. Clark

Prepared for
DIRECTORATE OF PLANNING AND TECHNOLOGY
DEVELOPMENT ENGINEERING DIVISION
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

Sponsored by
Advanced Research Projects Agency
Project VELA UNIFORM
ARPA Order No. 800

Project 6050
Prepared by
THE MITRE CORPORATION
Bedford, Massachusetts
Contract AF19(628)-5165

FOREWORD

The work reported in this document was performed under the Large Aperture Seismic Array (LASA) Project by the D-50 Department, The MITRE Corporation, Bedford, Massachusetts. The cognizant Air Force agency was the LASA Project Office, Directorate of Planning and Technology, Electronic Systems Division, L. G. Hanscom Field, Bedford, Massachusetts. This research is a part of Project Vela Uniform, which is sponsored by the Advanced Research Projects Agency of the Department of Defense; it is supported under Air Force Contract AF 19(628)-5165, (ARPA Order 800), Program Number 6250601R, and Systems Program 920F, LASA-ARPA.

REVIEW AND APPROVAL

Publication of this technical report does not constitute Air Force approval of the report's findings or conclusions. It is published only for the exchange and stimulation of ideas.

KENNETH B. FESS, Colonel, USAF
Chief, Development Engineering Division
Directorate of Planning and Technology

ABSTRACT

This report is intended as an introduction to one possible approach to the seismic classification problem. It develops a very general classification model using automatic non-parametric learning based on limited data of known classification. The model accepts discriminants extracted from the seismogram and yields the probability that the input was due to an earthquake or an explosion. Thus, the discriminants are assumed to be available as inputs. Pattern recognition as used here is defined, the classification procedure is outlined, the adaptive estimation of joint probability-densities from a finite number of multi-dimensional vectors of known classification (the learning model) is discussed, a simplified flow diagram of the learning model is presented, and the selection of necessary control parameters is investigated.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

SECTION I

INTRODUCTION

A large number of seismic events are easily classified on the basis of single discriminants such as epicenter. However, all of the events of interest cannot be categorized in this manner and there remains a subset of events for which a higher degree of decision-making sophistication is required. It is these remaining events which are of interest in the following study. For these events, the backbone of the classification model must be based on measurements related to the seismogram.

It will be assumed throughout that all events of interest have been a priori detected. Thus, the input waveform to the classification model will be known to contain an event. The objective will then be to separate the events into the dichotomy of earthquake or nuclear explosion or into even finer categorizations.

The development of the classification model will be divided into two major components. These are (i) the selection of a set of discriminants which is capable of classifying the event and (ii) the development of a mathematical model utilizing these discriminants for accomplishing the classification. This report will emphasize the latter.

Accordingly, the recognition system to be considered here will accept an appropriately chosen set of discriminants as its input and yield as its output, in the simplest case, the probability that the event was an earthquake or an explosion. It will have the ability to utilize simultaneously discriminants taken from mixed domains such as time, frequency, and frequency-wave number and will accomplish classification using the concepts of automatic, non-parametric pattern recognition based on limited input data. Thus, the system will be concerned with methods of automatically establishing decision criteria for classifying events as members of one or another class, when the only information available about any class is that which is contained in a given finite set of samples (having unknown statistics) known to belong to the class.

The recognition system will be developed with two distinct modes of operation; a learning mode in which the system is exposed to a sequence of events, each labeled according to the class or category to which it belongs, and a recognition mode in which new unlabeled events are classified as members of one or another of these classes. During the learning mode, the system develops class-criteria from the labeled events submitted to it, and during the recognition mode it uses these criteria for classifying unlabeled events.

An event will be represented by an N-dimensional vector or point whose components are the values of the N measurable discriminants or parameters describing the event. Events belonging to the same category will be represented by points scattered throughout some region of N-dimensional space in accordance with an unknown (non-parametrically learned) N-dimensional probability distribution function. For the case of two discriminants and two classes, the hypothetical two-dimensional probability densities generated from a limited set of samples are shown in Figure 1.

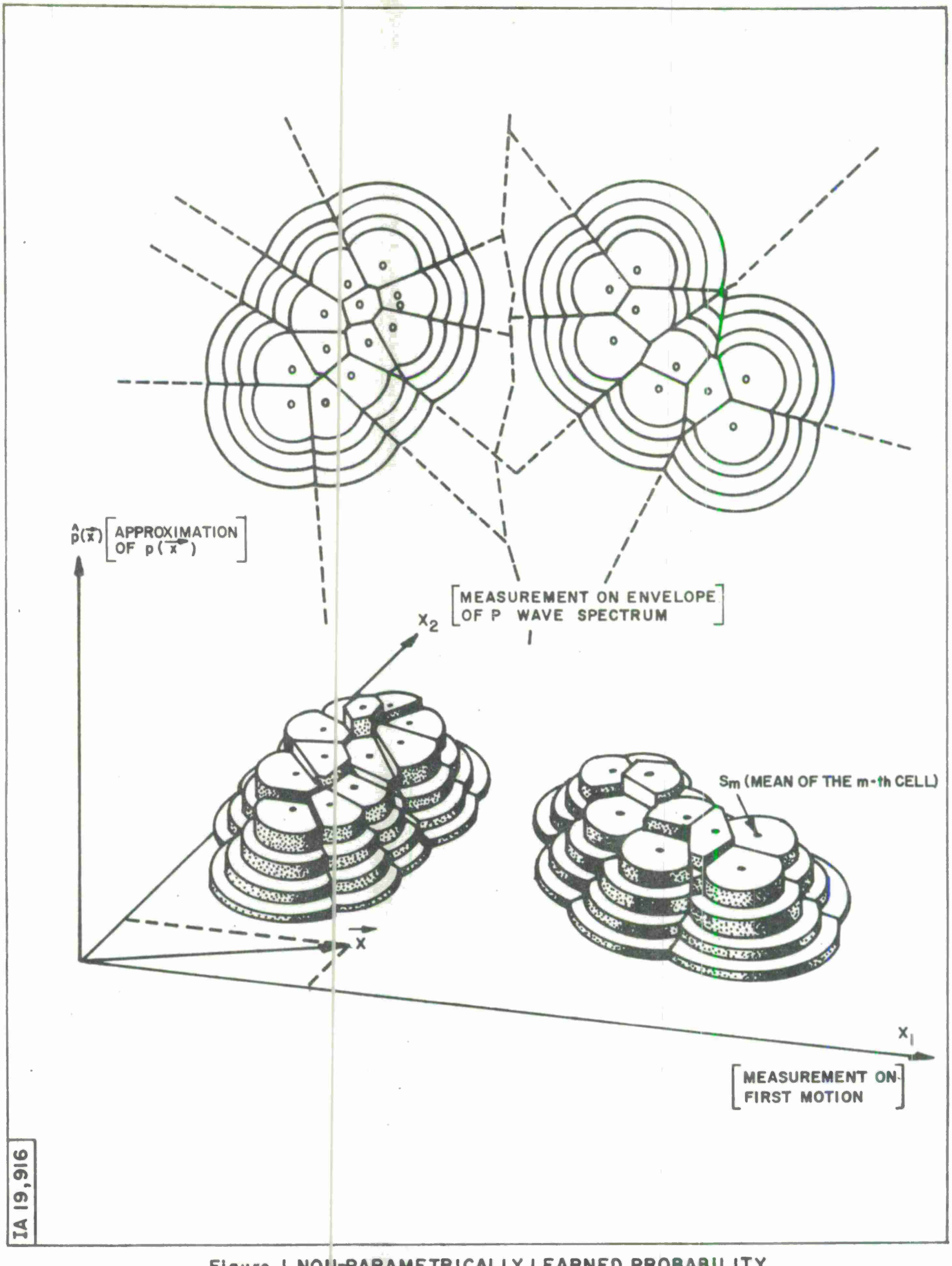
$\hat{p}(\vec{x})$ [APPROXIMATION OF $p(\vec{x})$]
x_2
[MEASUREMENT ON ENVELOPE OF P WAVE SPECTRUM]
S_m (MEAN OF THE m-th CELL)
$\vec{x}$
x_1
[MEASUREMENT ON FIRST MOTION]
IA 19,916

Figure 1 NON-PARAMETRICALLY LEARNED PROBABILITY DENSITY APPROXIMATION BASED ON LIMITED INPUT DATA FOR A FUNCTION OF TWO VARIABLES

SECTION II

PRELIMINARY DISCUSSION

Learning and recognition problems of pattern recognition can be formulated in mathematical terms as problems of recognition of membership in classes. The starting point of this method is to represent an input (in our case the seismic signal) by a set of measurements, variously called discriminants, clues, features, receptors, parameters, coordinate dimensions, properties or attributes. Accordingly, in this report, the terms clues or discriminants will be used interchangeably to describe measurements made on the time, frequency, frequency-wave number, etc., representation of the seismic signal. Each input that belongs to a given class (explosion, earthquake, etc.) will be regarded as a vector in a vector space which is located at a point defined by the discriminants. The class will then be represented by the collection of these points scattered in some manner in the vector space (often referred to as an observation or measurement space).

Members of different classes are distributed, in general, in different manners in the space. Machine learning (i.e. learning the pattern) is regarded as the problem of determining the best shape and location (i.e. best partitioning) of regions in the vector space so that the classes are distinguishable. This is illustrated in Figure 2. Pattern recognition or classification is the act of naming the region in which measurements made on an unknown seismic input are contained.

The three major parts of the pattern recognition system to be used here and their relationship to each other are illustrated by the block diagram of Figure 3. This shows the observation system that represents the seismic input by a set of measurements on this input or its transformations (discriminants). The choice of these

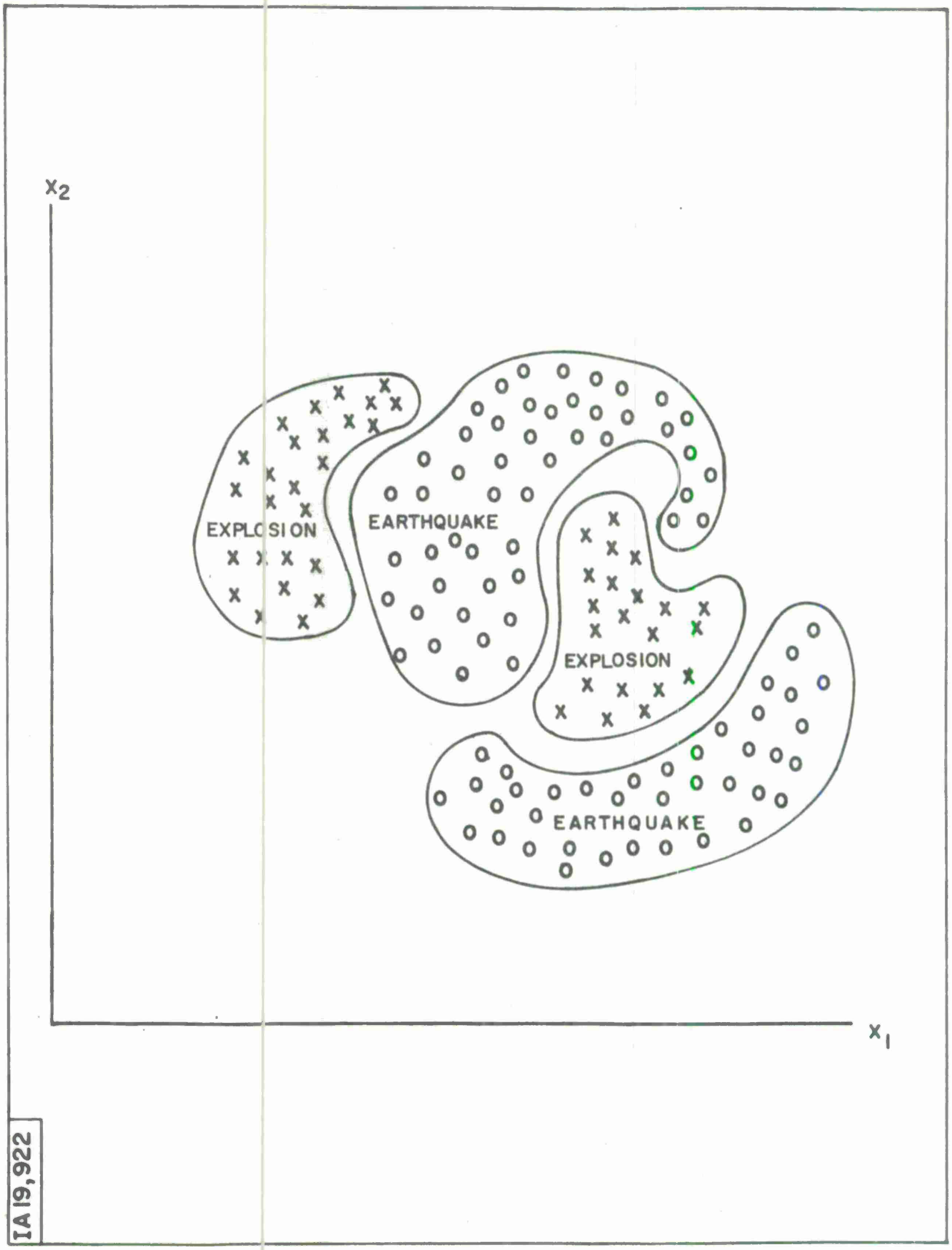

Figure 2. PARTITIONING OF VECTOR SPACE INTO REGIONS

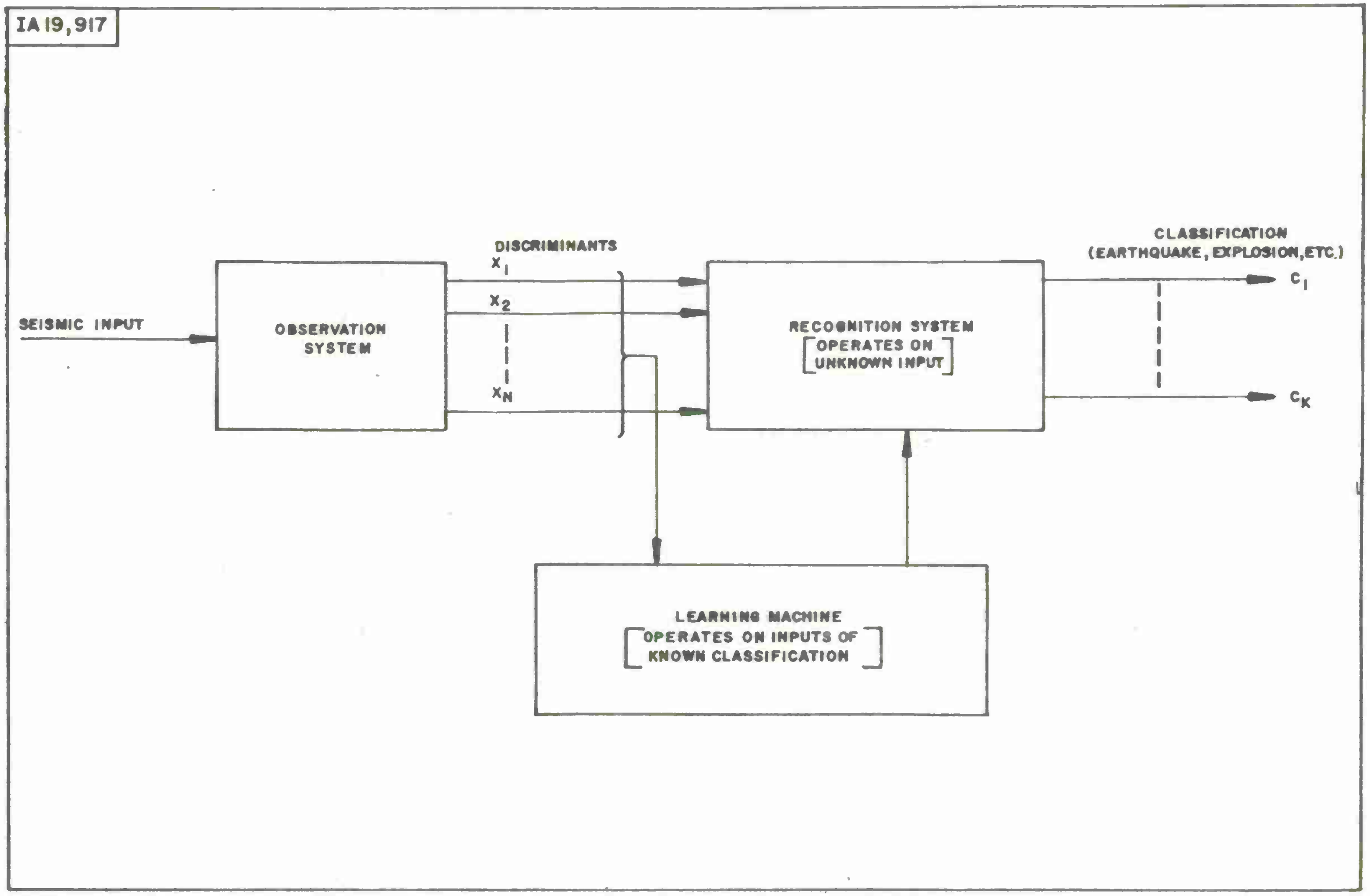
IA 19,917
SEISMIC INPUT
OBSERVATION SYSTEM
DISCRIMINANTS
X_1
X_2
X_N
RECOGNITION SYSTEM [OPERATES ON UNKNOWN INPUT]
CLASSIFICATION (EARTHQUAKE, EXPLOSION, ETC.)
C_1
C_K
LEARNING MACHINE [OPERATES ON INPUTS OF KNOWN CLASSIFICATION]

Figure 3 GENERAL PATTERN RECOGNITION SYSTEM

discriminants is an important problem which is presently being studied. It shows the "learning machine" in which seismic inputs of known classification are processed (for developing a good partition of the vector space). And, it shows the classification or recognition system which evaluates an unknown seismic input to decide in which partition of the space it is contained.

There are many ways of partitioning the vector space into regions. However, statistical methods (in particular, statistical decision theory) seem to be a leading contender for affecting good partitions. The applicability of decision theory in the design of pattern recognition systems is readily appreciated by considering its basic characteristics. Once input seismic stimuli are expressed in terms of a set of discriminants, we want to design a classification system with the best performance; i.e., one that makes the least number of mistakes. In addition, we recognize that the classification system will have to render decisions on inputs that are not identical to those from which classification was learned (although they will be similar, in general). It is well known that if we wish to minimize the risk, the probability of error, or the maximum error due to the decision we make, then we should make our decision by comparing likelihood ratios to fixed thresholds. That is to say, if we must choose between two classes, explosions and earthquakes, as giving rise to the seismic stimulus which we observe through a set of measurements on the seismic waveform or its transformations (discriminants), then the optimum decision is based on the comparison of the ratio of conditional probability densities with an appropriately chosen constant. In mathematical form, this expresses the notion that if the set of discriminants is a more likely occurrence under the assumption that the seismic stimulus belongs to the class of explosions than to the class of earthquakes, then common sense (and statistical techniques) advises us to decide that an explosion probably gave rise to our observations. Thus, decision theory provides us with a design procedure which reflects ultimate system performance as the basis for system design, and it also agrees with intuition.

There is a fundamental difference between the answers that are derivable from standard statistical techniques and the answers sought here. Usually, decision theory assumes knowledge of the relative frequency of occurrence of every observable set of discriminants from all classes of interest. Here, this state of knowledge is missing and estimates of the required quantities will be automatically made from a finite number of class samples. Thus, we recognize the fact that sparse seismic data with unknown statistics may be available and we design our system to account for this.

SECTION III

CLASSIFICATION BY LIKELIHOOD FUNCTION ESTIMATION

Consider the problem of deciding which of M classes has given rise to an observed event, $\vec{x} = (x_1, x_2, \ldots, x_N)$, and suppose that the statistics of events and classes are known, i.e., the joint probability density function of $\vec{x}$ and m is known, where m denotes the class label $(m = 1, 2, \ldots, M)$. The decision theoretical optimum method for processing a measured event $\vec{x}$ to render the classification is well known. Specifically, $\vec{x}$ should be regarded as a member of the k-th class if the cost of deciding in favor of the k-th class is less than that of deciding in favor of any of the other classes. This is stated in Equation 1.

$$\sum_{m=1}^{M} P_m p(\vec{x}|m) \left[C_K^{(m)} - C_z^{(m)} \right] \leq 0 \quad \text{for all} \quad z \neq k,\ z = 1, 2, \ldots, \qquad (1)$$

where

$C_z^{(m)}$ = the cost associated with deciding that $\vec{x}$ belongs to the z-th class when in fact $\vec{x}$ belongs to the m-th class,

P_m = the *a priori* probability that an event from class m will occur, and

$p(\vec{x}|m)$ = the conditional probability density functions of $\vec{x}$, given that $\vec{x}$ belongs to the m-th class.

This method of decision-making minimizes the average risk associated with the classifications. If, as is appropriate with many practical classification problems, the cost is the same for all misclassifications, then Equation 1 reduces to the following decision rule: decide $\vec{x}$ is a member of the k-th class if

$$P_k p(\vec{x}|k) \geq P_z p(\vec{x}|z) \quad \text{for all} \quad z \neq k,\ z = 1,\ 2,\ \ldots,\ m\ . \qquad (2)$$

Further, if the a priori probabilities are the same for all classes ($P_m = 1/M$ for all m), then Equation 2 becomes the following: decide $\vec{x}$ is a member of the k-th class if

$$L_{\vec{x}}(k) \geq L_{\vec{x}}(z) \quad \text{for all} \quad z \neq k,\ z = 1,\ 2,\ \ldots,\ m\ , \qquad (3)$$

where $L_{\vec{x}}(m) = p(\vec{x}|m)$ is commonly called the likelihood function of m given the event, $\vec{x}$. When class a priori probabilities are the same the likelihood function is equal to the a posteriori probability of class occurrence; i.e., $L_{\vec{x}}(m) = p(\vec{x}|m) = p(m|\vec{x})$.

In the event that H of the dimensions of the N-dimensional input vector $\vec{x}$ are not available (this corresponds to the case where H of the selected set of N discriminants cannot be extracted from the seismogram) and the classifier has been designed to operate as an N-parameter processor, it is not obvious what the optimum classification decision based on the N-H observed measurements consists of. However, a study has been carried out in the appendix for determining the method of making optimum decisions in this case. It is concluded from this study that the optimum decision based on N-H observed measurements consists of comparing the ratio of a posteriori probabilities of the actually observed N-H measurements with a threshold and that no useful purpose is served by knowing to what higher-dimensional process the N-H measurements belong. This agrees with intuition.

Thus, we see that if the statistics of the events in classes are known, then an optimum (from the standpoint of minimizing risk) method of establishing classification decision boundaries in observation space is known, and the only hurdle which remains is implementation of this procedure. Unfortunately, however, this result can only be used as a guide to solving the seismic classification problem, because the statistics of the seismic input are usually not known precisely.

In particular, in seismic classification, all of the information available on the statistics of the seismic signal is contained in the values of a finite number, N , of labeled samples from each of the M classes or categories. However, one can still proceed in this situation by generating estimates, using the available data samples, of the likelihood functions (or equivalently, the probability density functions) of the different classes over the observation space, and rendering classification decisions in a manner dictated by decision theory using the estimated quantities in lieu of the "true" functions. This is the basis for the classification method to be discussed in this report.

The process of estimating the probability densities from labeled samples of known classification can be regarded as "learning", while the evaluation of likelihood ratios according to optimum decision theory is called "recognition".

Some parametric learning methods assume that the functional form of the densities is known (except for a set of undetermined parameters), while non-parametric learning methods deliberately assume no knowledge of the form of the densities (although some assumptions of their "well-behaved" character is implicit). Because of the complicated nature and uncertainty of the form of the conditional joint probabilities involved in seismic signal-processing, expression of the densities in analytical form does not seem to be a reasonable classification solution. Instead non-parametric methods seem to be a more realistic approach.

Thus, the complexity of the seismic problem leads us to consider adaptive (non-parametric) methods of estimating the unknown (multimodal) densities. In particular, classification might be accomplished by storing non-parametrically determined values of the densities to be estimated at a sufficiently large number of points of the vector space, determining the stored point nearest to the unknown input $\vec{x}$, looking up the value of the density at the nearest point or perhaps, interpolating among stored values of the densities near $\vec{x}$, and rendering a classification decision based on the value of the observed density. This can be visualized in the one-dimensional case as shown in Figure 4. Note that fewer points can be used to represent the density in the region where the density does not vary much, and more points can be used where the density varies rapidly. Here the one-dimensional probability density $p(x_i)$ is approximated with a staircase approximation $\hat{p}(x_i)$. Similarly, an N-dimensional density involving the joint probability of occurrence of N different numerical values can be approximated by the N-dimensional equivalent of a staircase approximation. Such an approximation of a probability density is a histogram in N dimensions.

Since the density function $p(\vec{x})$ is approximated by a constant in each interval, it is obvious that only boundaries of the intervals and the values of the approximation must be stored. A simple method of evaluating a histogram approximation at an arbitrary point can thus be devised. The procedure hinges on the ability to determine simply the identity of the cell or interval, ν, in which the input to be classified is contained and then retrieving p_ν, the corresponding stored value of this approximation.

By storing the location of the centers of the cells as a set of points, $\{\vec{S}_\nu\}$, where $\vec{S}_\nu$ is the stored center of the ν-th cell, the interior of an arbitrary cell ℓ is readily defined as the locus of points "nearer" to $\vec{S}_\ell$ than any other stored point. The classification procedure thus implied is:

1. Determine the stored point $\vec{S}_\ell$ that is "nearer" to the input vector $\vec{x}$ than to any other point $\vec{S}_\nu$ (ν not equal ℓ).
2. Retrieve the stored probability density $p(\vec{S}_\ell)$ (approximately equal to $p(\vec{x})$) to estimate $p(\vec{x})$.

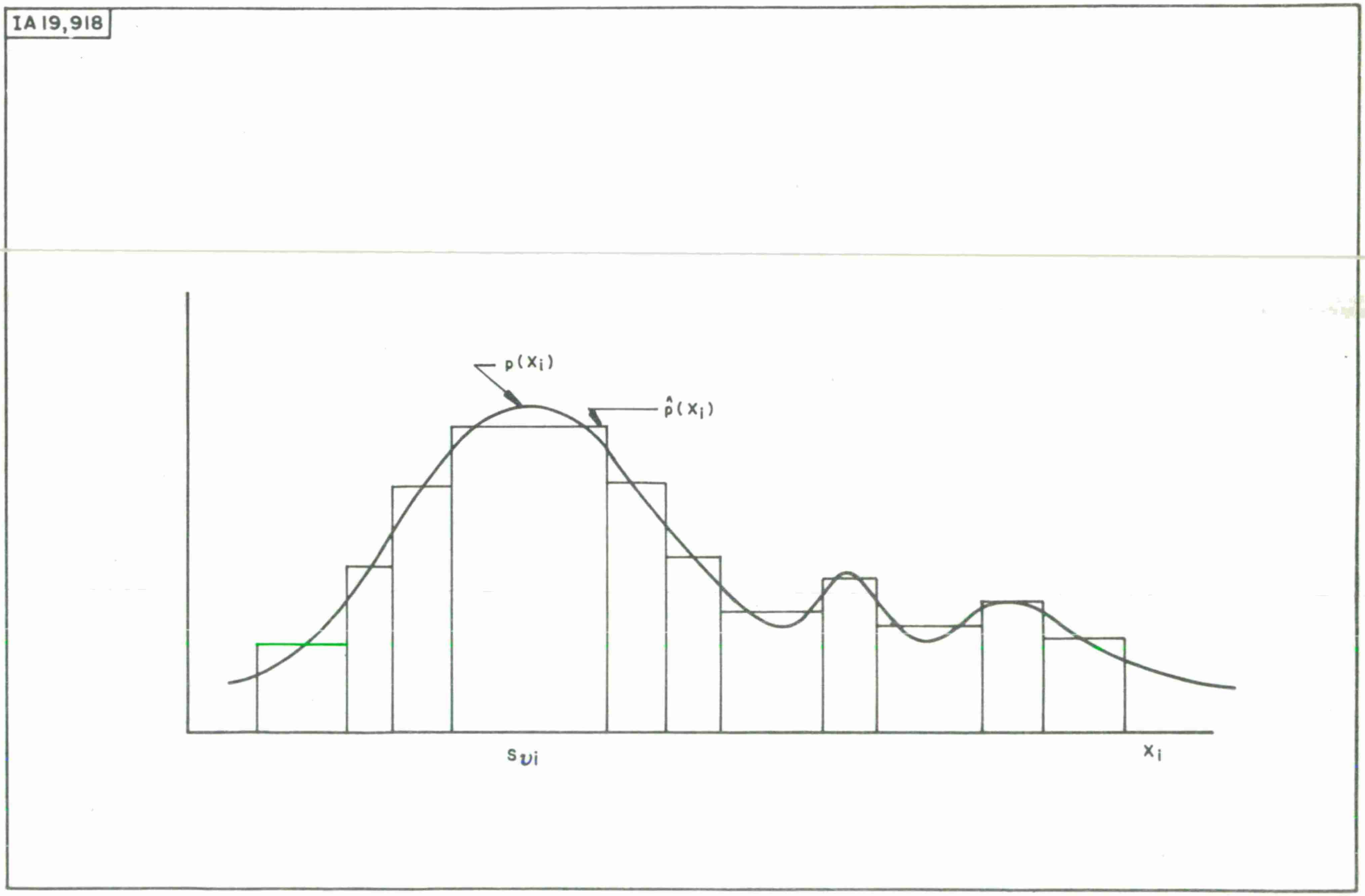

Figure 4. HISTOGRAM ESTIMATION USING UNEQUAL CELL SIZES

3. Repeat this procedure for all classes and compute the likelihood ratios, joint probabilities, etc. necessary for classification.

One can place this construction of histograms with unequal cell sizes on an exact mathematical basis by asking (and solving) questions of the type "What is the optimum choice for the location, size and height of the cells to minimize the expected error between $p(\vec{x})$ and its estimate, $\hat{p}(\vec{x})$?" However, since in practice $p(\vec{x})$ is unknown and must be obtained from samples, this is the same as attacking the problem of how to obtain a good histogram directly from the samples. It is readily appreciated that cells representing the distribution of a set of known samples of class k must be located in those regions in the vector space where members of class k are observed. Thus, it seems desirable to have members of the class create and determine the locations and dimensions of the histogram cells. Since the cell centers thus obtained typify the distribution of class k, the stored points $\{\vec{S}_\nu\}$ will be called "typical samples" of the class.

Since the interior of an arbitrary cell ℓ in the histogram is defined as the locus of points "nearer" to $\vec{S}_\ell$ than to any other stored point $\vec{S}_\nu$ (ν not equal to ℓ) one should postulate distance-measures which stretch when they measure "nearness" to a stored point $\vec{S}_\nu$ whose cell is wide, and shrink for a narrow cell. A squared distance measure exhibiting this property is expressed by the quadratic form $Q_\nu(\vec{x})$ given by

$$Q_\nu(\vec{x}) = \sum_{i=1}^{N}\left(\frac{x_i - S_{\nu i}}{\sigma_{\nu i}}\right)^2 , \qquad (4)$$

where ν identifies the cell and i is the specific dimension of the space under consideration. This quadratic form expresses the notion that the approximated density varies differently in one cell than in another, and it also expressed coordinate-direction-dependent differences in the rate of variation of the function. It is an expression of the location and also the shape of the cells of the N-dimensional histogram. Thus, a difference between parameter values of the input $\vec{x}$ and the stored sample $\vec{S_{\ell}}$ may be judged more significant in one neighborhood of the vector space than in another.

In the event that a new input vector does not fall within any cell, it will be assumed that the probability density is well behaved and exhibits Gaussian decay in regions where the probability density is small.

SECTION IV

THE ADAPTIVE APPROXIMATION OF PROBABILITY DENSITIES FROM LIMITED DATA

In the method of evaluation of probability densities described above, the approximated density was described by a set of typical samples and cell shapes determined by quadratic forms specified by means $\{S_{\nu i}\}$ and variances $\{\sigma^2_{\nu i}\}$. In the following, an algorithm is described for generating the cells from data in an adaptive manner by accepting input samples of known classification sequentially. A simplified flow chart illustrating the procedure is shown in Figure 5.

When the first learning sample is introduced, a cell of prechosen size and shape is created and is centered on the first learning sample. The initial size and shape of the cell is determined by prior analysis of the data (to be discussed in the next section) as part of the initializing procedure. The interior of the cell is defined by Equation 5, the equation of an ellipsoid in N dimensions, where the squared radii of the ellipsoid are expressed by $\sigma^2_{\nu i}(t)$, and τ^2_N is a threshold control parameter. In Equation 5, the symbol t signifies the fact that the cell center and shape are functions of the number of learning samples contained in the n-th cell up to the present time. T will denote the total number of inputs to the present.

$$Q_\nu(\vec{x}, t) = \sum_{i=1}^{N} \left(\frac{x_i - S_{\nu i}(t)}{\sigma_{\nu i}(t)} \right)^2 \leq \tau^2_N \qquad (5)$$

The first input vector becomes the first typical sample. This plus an estimate of the density, given by Equation 6, is stored. The density is estimated by the ratio of the fraction of the total number of input vectors that fall in a cell to the volume of that cell.

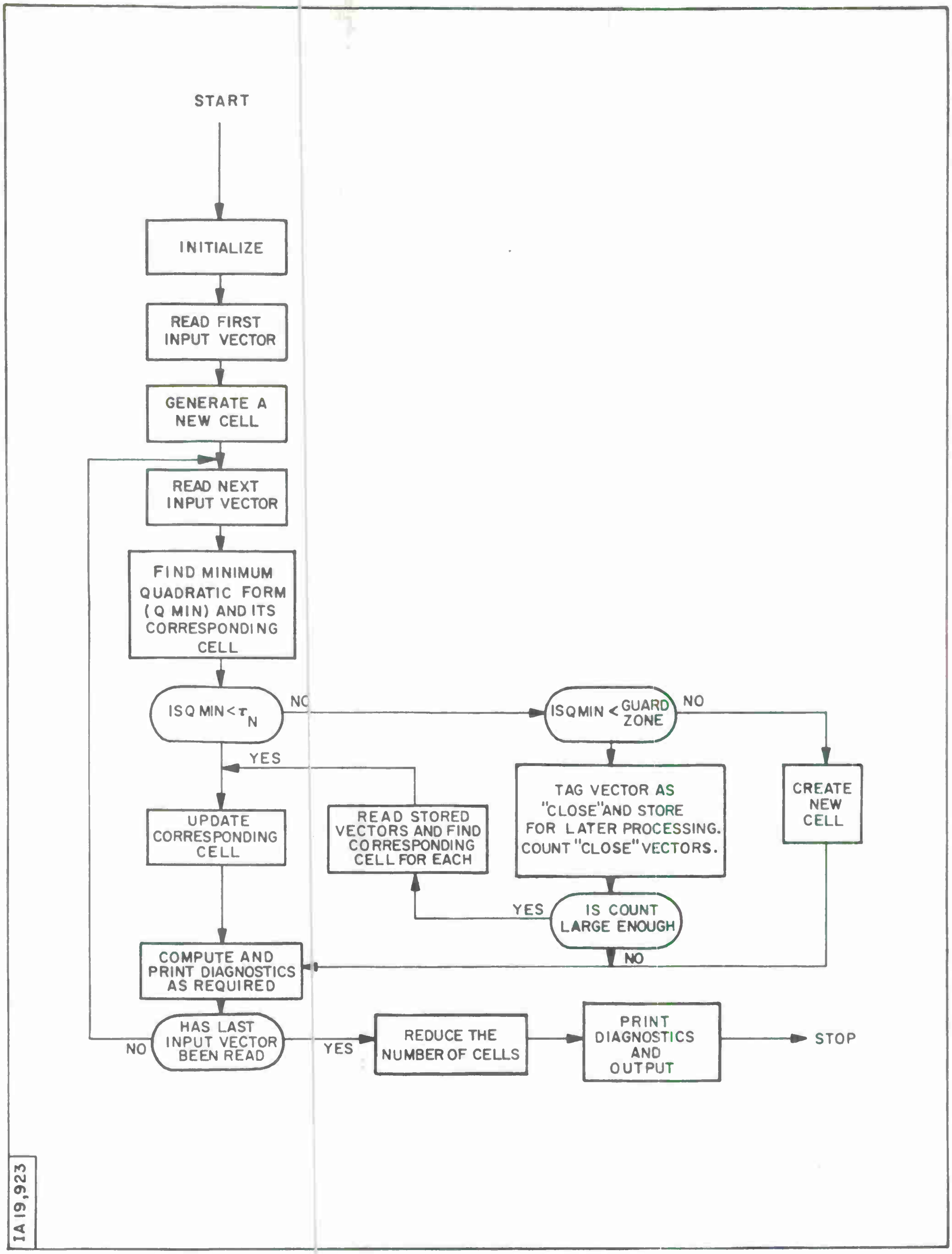
START
INITIALIZE
READ FIRST INPUT VECTOR
GENERATE A NEW CELL
READ NEXT INPUT VECTOR
FIND MINIMUM QUADRATIC FORM (Q MIN) AND ITS CORRESPONDING CELL
IS Q MIN < τ_N
NO
YES
UPDATE CORRESPONDING CELL
IS Q MIN < GUARD ZONE
NO
TAG VECTOR AS "CLOSE" AND STORE FOR LATER PROCESSING. COUNT "CLOSE" VECTORS.
CREATE NEW CELL
READ STORED VECTORS AND FIND CORRESPONDING CELL FOR EACH
YES
IS COUNT LARGE ENOUGH
NO
COMPUTE AND PRINT DIAGNOSTICS AS REQUIRED
NO
HAS LAST INPUT VECTOR BEEN READ
YES
REDUCE THE NUMBER OF CELLS
PRINT DIAGNOSTICS AND OUTPUT
STOP
IA 19,923

Figure 5. SIMPLIFIED FLOW DIAGRAM FOR A NON-PARAMETRIC LEARNING PROGRAM

Except for a constant k_N (given in the next section) that depends on the number of dimensions, the volume of the cell is expressed by the product of the standard deviations in the quadratic form used to define the boundaries of the cell.

$$p(\vec{S}_\nu, t) \approx \frac{t}{T}\left[\prod_{i=1}^{N} \sigma_{\nu i}(t)\right]^{-1} \tag{6}$$

The second learning vector is used to generate a second cell, similar to the first, if it falls sufficiently outside the first cell. However, if the second vector falls inside the first cell, the center of that cell is shifted to the mean of the two learning vectors, the shape and size of the cell is adapted from a better knowledge of the local distribution of members of the class, and the local estimate of the probability density is updated accordingly. If the second vector falls outside the first cell, but by not a large amount, it is stored temporarily to be reused at a later time according to a procedure to be described in subsequent paragraphs.

The third and subsequent learning vectors are processed similarly, either generating new cells, updating old cells, or being stored temporarily for later use. The cells so generated for each class are located only in the portion of the vector space where members of the individual classes have been observed.

It is seen through the above discussion that as learning vectors are introduced sequentially, the cell in the immediate neighborhood of the input vector changes shape, size, location and height. It is therefore important to examine the time-dependency of these cell parameters. Accordingly, the variances that determine the cell shape are given by Equations 7 and 8.

$$\sigma^2_{\nu i}(t) = \max\left[\sigma^2_{\nu i}(0),\ \xi_{\nu i}(t)\right] , \tag{7}$$

$$\xi_{\nu i}(t) = \frac{1}{t} \sum_{r=1}^{t} \left[x_{\nu i}(r) - S_{\nu i}(t)\right]^2 , \tag{8}$$

where

t denotes the number of input vectors that fell in the ν-th cell up to the present time,

$x_{\nu i}(r)$ is the i-th coordinate value of the r-th input vector falling in the ν-th cell,

$S_{\nu i}(t)$ is the i-th coordinate value of the n-th cell center after t contributions to the cell.

Equation 7 expresses the manner in which the i-th coordinate of the ν-th radius $\tau_N \sigma_{\nu i}(t)$ grows if the sample variance $\xi_{\nu i}(t)$ of the t vectors in the cell exceeds the initial variance $\sigma^2_{\nu i}(0)$. The cell radius is never allowed to shrink to less than the initial value $\tau_N \sigma_{\nu i}(0)$. The reason for defining the cell in this way is to encourage the cells to increase in size as more inputs are received, thus keeping the total number of cells used in the approximation of the probability density small.

To insure that each cell can grow while reducing the chance for an overlapping coverage of the same region of the vector space by several cells, an outer control parameter $(\theta \geq 1)$ is introduced. Thus, a vector $\vec{x}$ not falling within an existing cell (as defined by the threshold τ_N) is used to generate a new cell only if it is outside the larger concentric cell defined by Equation 9.

$$Q_\nu(\vec{x}, t) \leq (\theta \tau_N)^2 \tag{9}$$

It is seen that the quantity θ expresses the ratio of the outer to inner diameter of a "guard zone" within which input vectors neither create new cells nor update old ones.

The input vectors which neither create or update cells are stored temporarily for later use. As the cells grow in size, these stored vectors can be forced into the existing cell structure without the need to create new cells.

After a cell structure is obtained by the procedure described above, we may find that the number of cells created is larger than the number we would like to have in the N-dimensional generalized histogram. We may force the reduction of the number of cells created by altering the cell growth controlling parameters τ_N and θ. In most cases, however, it can be expected that a significant percentage of the cells created will contain very few input vectors, and, in general, these sparcely populated cells will surround the more populous cells. This will happen because each cell center (typical sample), after the cells initial creation, will migrate in the vector space and tend toward the nearest mode (local peak) of probability density to be approximated. This is readily seen from the one-dimensional illustration shown in Figure 6.

This figure shows a small range of the variable x_i and the probability density $p(x_i)$ in that interval. The point $S_{\nu i}(t)$ represents the cell center of i-th coordinate of the ν-th cell after t members fell into the cell. The probability is greater that the next input is to the right of $S_{\nu i}(t)$ than that it is to the left of that point. This implies that the cell center will move to the right after the t-plus-first input falling within the ν-th cell is introduced. It is thus seen that cells migrate in the direction of the nearest modes. As cells move toward modes, and later inputs create cells at places from which older cells have migrated, there will always be some cells which contain few members. Thus the number of cells can be reduced by forcing cell locations containing few members into the nearest cells whose members exceed a predetermined number.

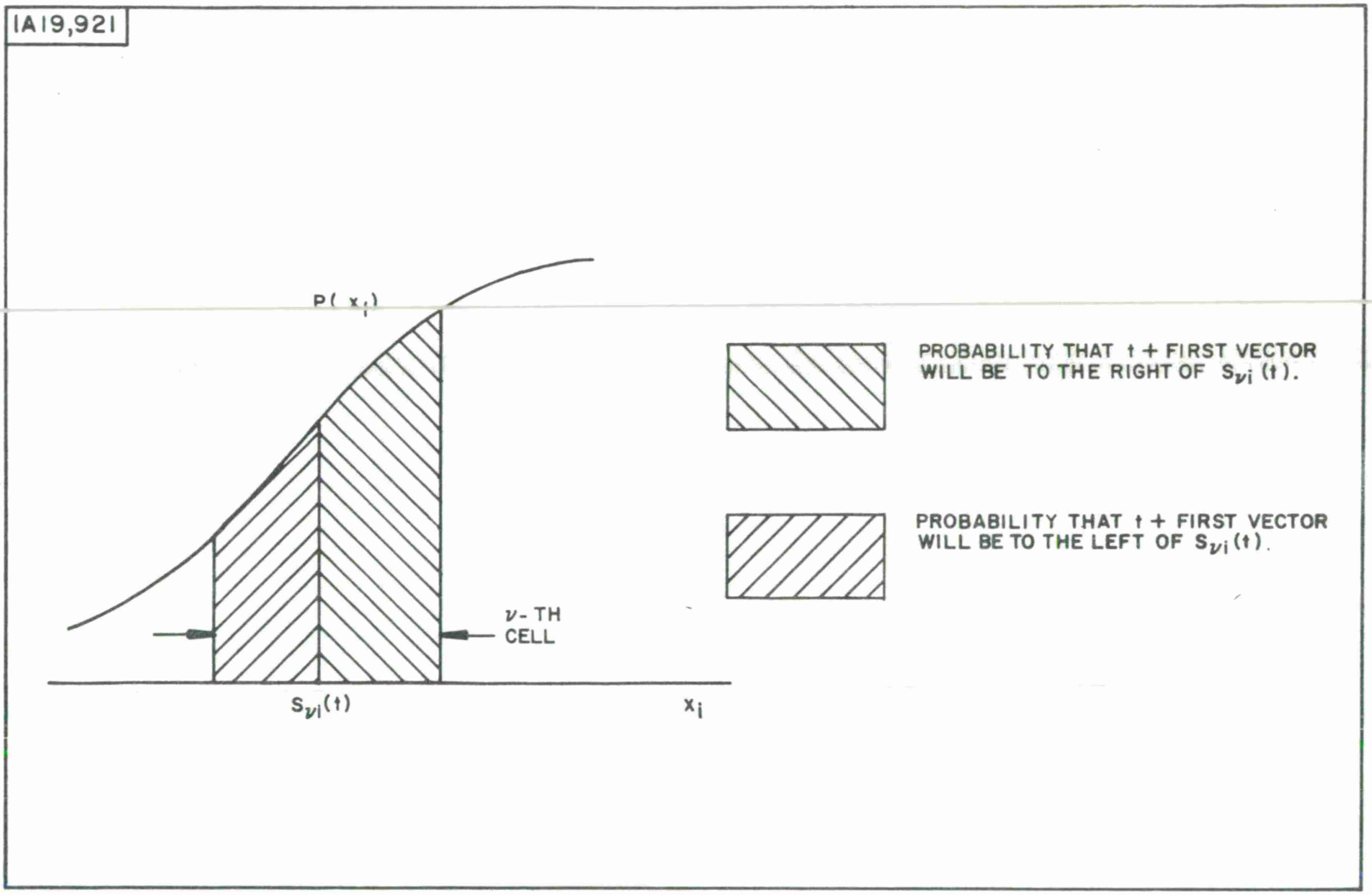
IA19,921
P(x_i)
PROBABILITY THAT t + FIRST VECTOR WILL BE TO THE RIGHT OF $S_{\nu i}$ (t).
PROBABILITY THAT t + FIRST VECTOR WILL BE TO THE LEFT OF $S_{\nu i}$(t).
ν- TH CELL
$S_{\nu i}$(t)
x_i

Figure 6. MODE SEEKING PROPERTY OF CELLS

SECTION V

CONTROL PARAMETER SELECTION THEORY

In the following paragraphs, some of the properties of the cell growth mechanism will be discussed. It is desirable that the individual cells be adjusted by the data so that a good approximation to the class probability density function should be obtained with a minimum number of cells. Furthermore, the size and shape of the individual cells should be determined by a reasonable and automatic procedure from the data in order to relieve the experimenter from the almost impossible task of picking appropriate cell sizes.

For simplicity, consider an isolated cell ν and let $\vec{x}(t)$ be the t-th observation point (known class member) that falls in the cell, let $\overrightarrow{S_\nu(t)}$ be the sample mean of the first t observations that lie in the cell (i.e. $\overrightarrow{S_\nu(t)}$ is the center of the cell at the t-th step), let $\overrightarrow{\sigma_\nu(t)}$ be a vector weighting parameter determined according to Equations 7 and 8, indicating the cell shape, and finally, let τ_N be a scalar constant (the constant τ_N is the control parameter being studied here). Then, the cell is defined at the t-th step to be the set of points in the observation space defined by Equation 5 (repeated below for convenience).

$$Q_\nu(\vec{x}, t) = \sum_{i=1}^{N} \left(\frac{x_i - S_{\nu i}(t)}{\sigma_{\nu i}(t)} \right)^2 \leq \tau_N^2 \qquad (5)$$

Thus, the cell is the (ellipsoidal) locus of points "closer" to the cell mean $S_{\nu i}(t)$ than $\tau_N \sigma_{\nu i}(t)$ in the i-th direction. It should be emphasized again that such a cell is "mode seeking" in that it will move (as a function of t) in the direction of the greatest concentration of data points. This is a very desirable feature. The cell is first established according to some rule by a data point which does not fall in any other cell so that $\overrightarrow{S_\nu}(1) = \vec{x}(1)$, i.e., the cell is initially centered about the first or establishing data point. If $\overrightarrow{\sigma_\nu}(t) = \overrightarrow{\sigma_\nu}(0)$ for all t , the cell size and shape remains the same throughout the estimation process. Then the choice of $\overrightarrow{\sigma_\nu}(0)$, which is based largely on physical considerations and intuition, is very critical and an intelligent choice is very difficult. But, if $\overrightarrow{\sigma_\nu}(t)$ is made to depend on the data sample, the volume of the cell may be made to grow to an "optimum" size by proper choice of the constant τ_N . Although the cell might alternatively be made to shrink if the data indicated this were desirable, it is assumed here that the initial cell size is small compared to intervals in which the class probability density function changes greatly and, hence, only cell expansion is discussed below.

The rule for updating the vector $\overrightarrow{\sigma_\nu}(t) = \left[\sigma_{\nu 1}(t), \sigma_{\nu 2}(t), ---, \sigma_{\nu N}(t)\right]$ is found from Equations 7 and 8 as

$$\sigma^2_{\nu i}(t) = \max\left\{\sigma^2_{\nu i}(0),\ \xi_{\nu i}(t) = \frac{1}{t}\sum_{r=1}^{t}\left[x_{\nu i}(r) - S_{\nu i}(t)\right]^2\right\}. \qquad (10)$$

Thus, $\sigma_{\nu i}(t)$ begins at a preset value and normally grows to be the sample standard deviation of the sample vectors in the cell neighborhood.

The radius of the cell, defined by Equation 5 in the i-th coordinate direction, is $\sigma_{\nu i}(t)\tau_N$. The constant τ_N is chosen according to the theory to be developed here, however, the initial cell radii $\sigma_{\nu i}(0)\tau_N$ must still be selected on the basis of physical considerations.

The cell volume might be considered optimum if it is as large as possible and still yields an estimated probability density function consistent with that obtained by estimating over smaller cells. If a cell is located in a region of the observation space over which the class probability density function is a constant, the cell size should expand until it covers the region of uniform distribution. Furthermore, once the cell is "firmly established" in the sense that a number of observations have fallen in the cell, the rate of expansion should be fairly rapid provided it does not grow substantially beyond the region of constant probability density function. On the other hand, if the cell is initially located in a region over which the class probability density function is changing, the cell should not expand rapidly but should migrate toward a node. Therefore, the rule for updating $\overrightarrow{\sigma_\nu}(t)$ and the choice of τ_N should be such that the expected cell behavior obeys these two intuitive rules.

Using the above notions, we are now in a position to construct a model of the cell growth mechanism through a study of the random behavior of the cell. Accordingly, the volume of an N-dimensional ellipsoid is (the t and ν designators will be temporarily omitted for convenience)

$$V_N = k_N \prod_{i=1}^{N} \sigma_i \ , \tag{11}$$

when

$$\sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} = 1 \tag{12}$$

specifies the N-dimensional ellipsoid and

$$k_N = \frac{\pi^{N/2}}{\Gamma\left(\frac{N}{2}\right) + 1} \ . \tag{13}$$

A short table of k_N is given below

N	1	2	3	4	5	6	7	8	9
k_N	2	π	$\frac{4\pi}{3}$	$\frac{\pi^2}{2}$	$\frac{8\pi^2}{2}$	$\frac{\pi^3}{6}$	$\frac{16\pi^3}{105}$	$\frac{\pi^4}{24}$	$\frac{32\pi^4}{945}$

A slice perpendicular to the x_j-axis at x_j is an $(N-1)$-dimensional ellipsoid specified by

$$\sum_{\substack{i=1 \\ i\neq j}}^{N} \frac{x_i^2}{\sigma_i^2\left(1 - \frac{x_j^2}{\sigma_j^2}\right)} = 1 \;, \tag{14}$$

of volume

$$k_{N-1}\left(1 - \frac{x_j^2}{\sigma_j^2}\right)^{\frac{N-1}{2}} \prod_{i\neq j}\sigma_i \;. \tag{15}$$

Assuming a uniform probability distribution over the N-dimensional ellipsoid specified by Equation 12, the probability density function of the x_j coordinate is

$$g_N(x_j) = \frac{k_{N-1}\left(1 - \frac{x_j^2}{\sigma_j^2}\right)^{\frac{N-1}{2}} \prod_{i\neq j}\sigma_i}{k_N \prod_{i=1}^{N}\sigma_i} \;. \tag{16}$$

However, since the volume V_N can be obtained by integrating Equation 15 over x_j, it is found that

$$V_N = k_N \prod_{i=1}^{N} \sigma_i = 2 \int_0^{\sigma_j} k_{N-1} \left(1 - \frac{x_j^2}{\sigma_j^2}\right)^{\frac{N-1}{2}} \prod_{i \neq j} \sigma_i dx_j ,$$

or

$$V_N = k_{N-1} \beta\left(\frac{1}{2}, \frac{N+1}{2}\right) \prod_{i=1}^{N} \sigma_i , \qquad (17)$$

where

$$\beta\left(\frac{1}{2}, \frac{N+1}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{N}{2} + 1\right)} \qquad (18)$$

is the beta function. Accordingly,

$$g_N(x_j) = \frac{\Gamma\left(\frac{N}{2} + 1\right)}{\Gamma\left(\frac{N+1}{2}\right) \Gamma\left(\frac{1}{2}\right)} \frac{\left(\sigma_j^2 - x_j^2\right)^{\frac{N-1}{2}}}{} , \text{ if } -\sigma_j \leq x_j \leq \sigma_j , \qquad (19)$$

$$= 0 , \text{ if } |x_j| > \sigma_j .$$

Making the transformation of variables $x_j = \lambda_j \sigma_j$, the probability density function of λ_j is

$$h_N(\lambda_j) = \frac{\Gamma\left(\frac{N}{2}+1\right)}{\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \left(1 - \lambda_j^2\right)^{\frac{N-1}{2}} , \text{ for } |\lambda_j| \le 1 , \qquad (20)$$

$$= 0 , \text{ for } |\lambda_j| > 1 .$$

The maximum value of $g_N(x_j)$ is $k_{N-1}/k_N\sigma_j$ and of $h_N(\lambda_j)$ is k_{N-1}/k_N . Examples of $h_N(\lambda_j)$ for several values of N are shown in Figure 7. It will be shown later that for small λ_j , $h_N(\lambda_j)$ is approximately gaussian.

The mean and variance of λ_j are easily found to be

$$\overline{\lambda_j} = 0 , \qquad (21)$$

and

$$\text{Var } \lambda_j = \overline{\lambda_j^2} = \frac{k_{N-1}}{k_N} \int_{-1}^{1} \lambda_j^2 \left(1 - \lambda_j^2\right)^{\frac{N-1}{2}} d\lambda_j ,$$

$$= \frac{k_{N-1}}{k_N} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \text{Sin}^2\, \theta \text{ Cos}^N\, \theta d\theta ,$$

$$= \frac{k_{N-1}}{k_N} \left[- \frac{1}{N+2} \text{Sin } \theta \text{ Cos }^{N+1}\theta \Big|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} + \frac{1}{N+2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \text{Cos}^N\, \theta d\theta \right] ,$$

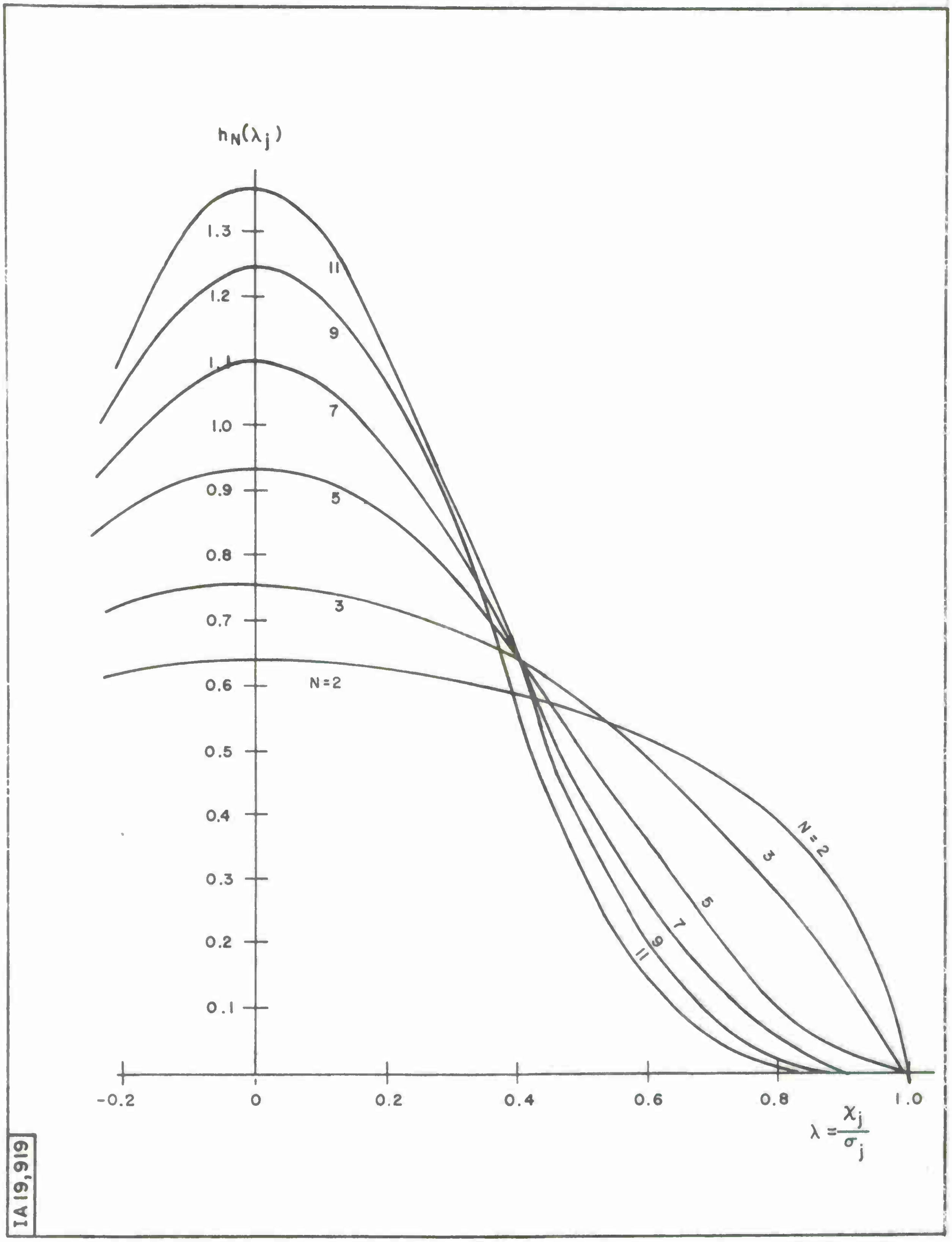

Figure 7. P.D.F. OF ONE COORDINATE OF POINTS UNIFORMLY DISTRIBUTED OVER AN ELLIPSOID OF N-DIMENSIONS

$$= \frac{k_{N-1}}{k_N} \; \frac{1}{N+2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{Cos}^N \, \theta d\theta \quad .$$

Thus,

$$\mathrm{Var} \; \lambda_j = \frac{1}{N+2} \, \frac{k_{N-1}}{k_N} \, \frac{\sqrt{\pi}\, \Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{N}{2} + 1\right)} \; ,$$

$$= \frac{1}{N+2} \quad . \tag{22}$$

To show that $h_N(\lambda_j)$ is approximately gaussian for small λ_j, rewrite Equation 20 as

$$h_N(\lambda_j) = \frac{\Gamma\left(\frac{N}{2} + 1\right)}{\sqrt{\frac{N-1}{2}}\, \Gamma\left(\frac{N+1}{2}\right)} \sqrt{\frac{N-1}{2\pi}} \left(1 - \lambda_j^2\right)^{\frac{N-1}{2}} \quad . \tag{23}$$

The first factor tends to unity as $N \rightarrow \infty$ by Stirling's formula. The logarithm of the third factor is

$$\frac{N-1}{2} \, \ell n \left(1 - \lambda_j^2\right) = \frac{N-1}{2} \left(- \lambda_j^2 - \frac{\lambda_j^4}{2} - \frac{\lambda_j^6}{3} - - - - \right),$$

$$= \frac{\lambda_j^2}{2 \frac{1}{N-1}} \left(1 + \frac{\lambda_j^2}{2} + \frac{\lambda_j^4}{3} + - - - \right) .$$

Hence, for any fixed λ_j such that $1 \gg \left(\frac{\lambda_j^2}{2} + \frac{\lambda_j^4}{3} + - - - \right)$,

$$\left(1 - \lambda_j^2\right)^{\frac{N-1}{2}} \cong \exp -\left[\frac{\lambda_j^2(N-1)}{2}\right] . \qquad (24)$$

Accordingly, in this region, $h_N(\lambda_j)$ is approximately normal with probability density function given by

$$h_N(\lambda_j) \cong \sqrt{\frac{N-1}{2\pi}} \quad \exp\left[-\frac{\lambda_j^2 (N-1)}{2}\right] . \qquad (25)$$

The mean of this distribution is zero and the variance is $1/N-1$. The tails of the distribution of λ_j decreases much faster than for the gaussian distribution (i.e. where λ_j is large) and $h_N(\lambda_j)$ goes to zero at ± 1 .

Letting

$$\delta_j(t) = \begin{cases} 0 , & t = 1 \\ X_j(t) - S_j(t-1) , & t > 1 , \end{cases} \qquad (26)$$

the t-th cell center is located at

$$S_j(t) = S_j(t-1) + \frac{1}{t}\,\delta_j(t) . \qquad (27)$$

Thus, substituting Equation 27 into Equation 8, the cell sample variance becomes (omitting the ν index but including the t index)

$$\xi_j(t) = \frac{1}{t}\left\{\sum_{r=1}^{t-1}\left[x_j(r) - S_j(t-1) - \frac{1}{t}\,\delta_j(t)\right]^2 + \left(\frac{t-1}{t}\,\delta_j(t)\right)^2\right\} ,$$

$$= \frac{1}{t}\left\{(t - 1)\ \xi_j(t - 1) + \frac{(t-1)}{t^2}\ \delta_j^2(t) + \left(\frac{t-1}{t}\ \delta_j(t)\right)^2\right\},$$

$$= \frac{t-1}{t}\left\{\xi_j(t - 1) + \frac{\delta_j^2(t)}{t}\right\},$$

$$= \frac{1}{t}\sum_{r=0}^{t-2} \frac{t-r-1}{t-r}\ \delta_j^2\ (t - r) \quad . \qquad (28)$$

Let $t = t'$ be the index of the first sample point for which $\sigma_j(t') > \sigma_j(0)$, i.e., the first time cell growth occurs. Then for $t - t'$, by Equations 5 and 22, the expected value of $\xi_j(t)$ is

$$\overline{\xi_j(t)} = \frac{1}{t}\sum_{r=0}^{t-2} \frac{t-r-1}{t-r}\ \overline{\delta_j^2(t - r)}$$

$$= \frac{\sigma_j^2(0)\tau_N^2}{N + 2}\ \frac{1}{t}\sum_{r=0}^{t-2} \frac{t-r-1}{t-r}\ , \quad t \le t' \ , \qquad (29)$$

or

$$\overline{\xi_j(t)} \rightarrow \frac{\sigma_j^2(0)\tau_N^2}{N + 2} \quad \text{as} \quad t \rightarrow \infty \ . \qquad (30)$$

From Equations 10 and 29 it is seen that a necessary condition for $\sigma_j(t)$ to be greater than $\sigma_j(0)$ so that cell growth may be expected to begin is

$$\overline{\xi_j(t)} \ge \sigma_j^2(0) \ ,$$

$$\frac{\sigma_j^2(0)\tau_N^2}{N + 2} \cdot \frac{1}{t} \sum_{r=0}^{t-2} \frac{t-r-1}{t-r} \geq \sigma_j^2(0) \quad ,$$

or

$$\tau_N^2 > N + 2 \quad , \tag{31}$$

since

$$\frac{1}{t} \sum_{r=0}^{t-2} \frac{t-r-1}{t-r} < 1 \quad ,$$

for $t < \infty$.

Furthermore, the choice of τ_N determines not only whether the cell may be expected to grow, but also the number t' of observations that must fall in the cell before cell growth can be expected to begin. It is desirable that t' be chosen sufficiently large to establish a firm cell location before the cell may be expected to grow. On the other hand, since the amount of data available for probability density function estimation is always limited in practice, t' must not be too large.

Having chosen an appropriate value for t' , the choice of the control parameter τ_N becomes automatic. Writing $\tau_N = \beta\sqrt{N + 2}$, and considering β as an unknown, Equation 29 can be solved for β .

$$\beta = \left[\frac{\overline{\xi_j(t)}}{\sigma_j^2(0)} \cdot \frac{1}{\frac{1}{t} \sum_{r=0}^{t-2} \frac{t-r-1}{t-r}} \right]^{\frac{1}{2}} , \; t \leq t \quad . \tag{32}$$

However, since t' is the index of the first sample point for which $\overline{\sigma_j^2(t')} > \sigma_j^2(0)$ or $\overline{\xi_j(t')} > \sigma_j^2(0)$, it is seen that for

$$\beta' = \left[\frac{1}{t'}\sum_{r=0}^{t'-2}\frac{t'-r-1}{t'-r}\right]^{-\frac{1}{2}} , \tag{33}$$

the choice of

$$\tau_N = \beta'\sqrt{N+2} \tag{34}$$

will result in a beginning of cell growth after an average of t' sample points fall in the cell.

A curve of β' as function of t' is shown in Figure 8. For the particular choice of $\tau_N = 1.4\sqrt{N+2}$, the curve in Figure 8 indicates that the average value of t will be approximately 4.7 before cell growth begins.

Since the probability density functions of greatest interest will more than likely be non-uniform over the entire space, there will, in general, be a wide spread in the range of cell probabilities. Therefore, the cells with high probabilities will normally begin to grow before the majority of the other cells have collected t' observations. Since the growth of an individual cell is limited by the presence of surrounding cells, it is reasonable to expect that in many instances the cells located near the modes of the distribution will have grown to their maximum limit by the time an average of t' points have been processed for each of the cells in the entire cell structure. This phenomemon requires further investigation.

An investigation of the dynamics of the growth mechanism should be carried out to shed more light on the method of selecting control parameters as discussed here. Experimentation should be of value in indicating if modifications to the above theory are necessary.

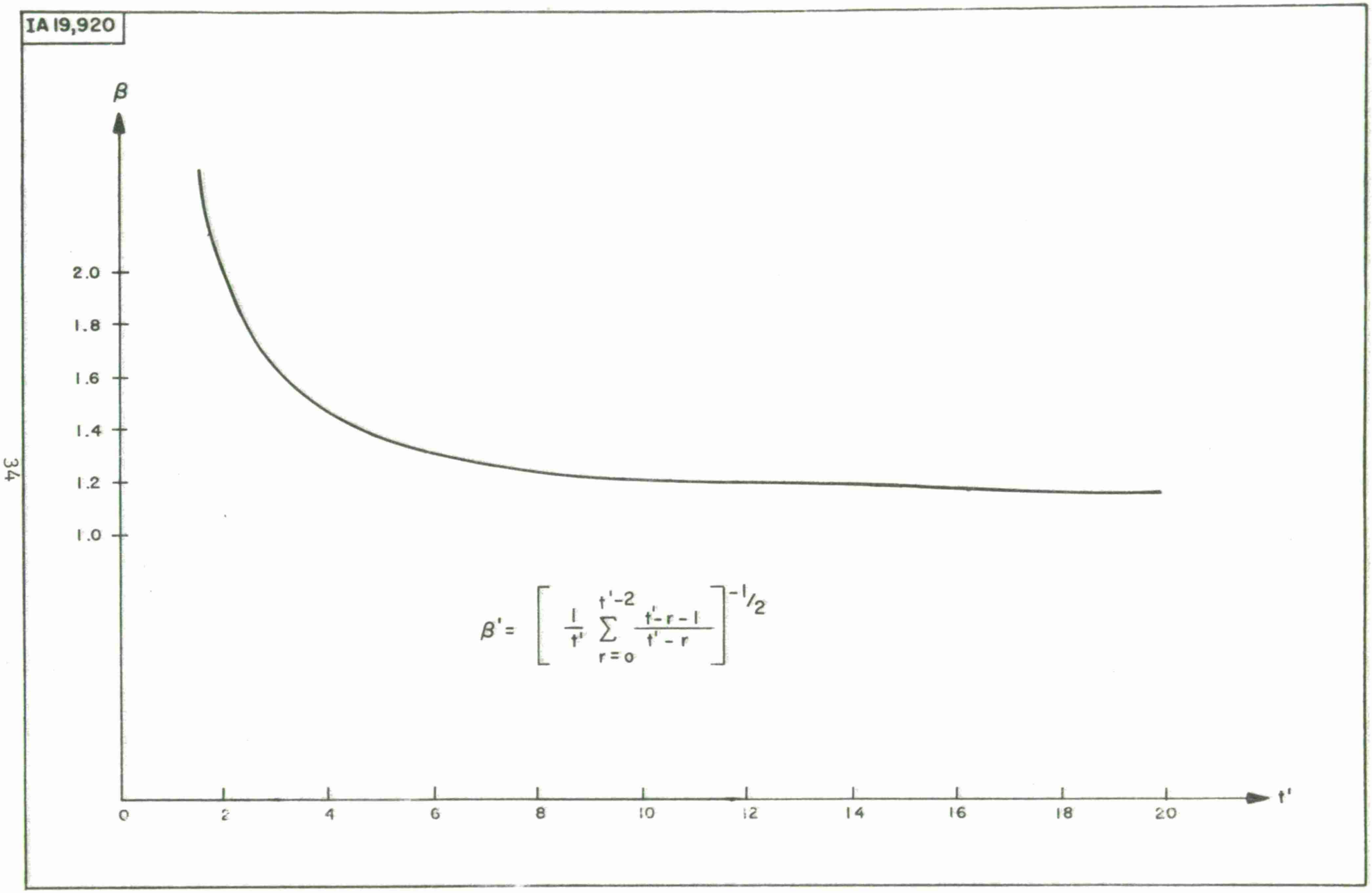

Figure 8. CURVE USED FOR SELECTION OF THE CONTROL PARAMETER τ_N

SECTION VI

SUMMARY

A very general classification model has been developed using the concepts of non-parametric pattern recognition based on limited data of known classification. The fundamental difference between decisions derivable from standard statistical techniques and decisions based on this model is that decision theory assumes knowledge of the relative frequency of occurrence of every observable set of discriminants from all classes of interest, while here, this knowledge is missing and estimates of the required statistical quantities are automatically made from a finite number of known class samples.

The model has two distinct modes of operation, a learning mode and a recognition mode. In the learning mode, partitioning of an N-dimensional parameter space (using discriminants derived from the seismic signal as coordinates) is accomplished by estimating the joint probability densities of the parameters for each of the input classes in question. In the recognition mode, maximum-likelihood ratio decisions on the estimated joint densities are made. It is significant that, in the learning mode, the estimates are formulated with cells which adjust their size automatically according to the data so that a good approximation to the class density function is obtained with a minimum number of cells. It is also significant that the cells are mode-seeking in that they move as new data is introduced in the direction of the greatest concentration of data points.

The entire development has been of necessity introductory, intended to give insight into the broad concepts involved. Thus, many of the problems of implementing the model and integrating it as a working part of a seismic signal processor require further theoretical studies as well as experimental verifications.

SECTION VII

RECOMMENDATIONS FOR FURTHER STUDY

Only a few of the many study problems which come to mind as a result of this study are listed below. In general, one is interested in developing clustering transformations, determining the quality of decisions rendered by the model, and gaining knowledge about the habits of the model's performance. In particular:

1. The cell growth mechanism should be studied theoretically in much more detail. For example, the effects of selecting $\sigma_{\nu i}(0)$ and θ should be explored as a function of the size of the data sample.
2. An experimental study of the control parameter selction problem should be undertaken.
3. The accuracy of probability density estimation should be experimentally determined using real and synthesized data.
4. The quality of the estimation procedure, both for the purpose of determining the reliability of the decision rendered in any one instance and for the purpose of modifying the learning procedure to yield decisions with lower error probabilities, should be theoretically investigated.
5. The proposed model should be compared to other techniques on the basis of error probabilities, complexity of implementation, etc.
6. Transformations on the original space should be considered for increasing apparent class separability. One such transformation might be concerned with minimizing entropy. Thus, for the density $p_i(y)$, one might minimize

$$H_i(y) = - \int p_i(y) \log p_i(y) \, dy ,$$

since $H_i(y)$ is a function only of the manner in which class members are distributed in observation space.

7. The order dependency of the probability density estimation technique should be investigated.

APPENDIX

CLASSIFICATION DECISIONS BASED ON INCOMPLETE SETS OF OBSERVATIONS

The problem of classifying signals is generally treated as a problem of optimally deciding, on the basis of N observed measurements on that signal $(x_1, x_2, \ldots x_N)$, which of several classes produced the particular set of N measurements. If the joint probability densities of the N measurements are known under all assumptions of class membership for the set of N measurements, classification decisions can be rendered by computing the N-dimensional likelihood ratios and then comparing these ratios with each other or with a threshold. With the joint probability densities not known, but a finite number of measurements of each class available, decision rules can be devised which approach the likelihood ratio computation in the limit, as the number of measurements approaches infinity.

The problem which will be considered in this appendix concerns the method of making the optimum decision when not all of the N-measurable parameters of an N-dimensional process are available. It will be assumed for simplicity that the incomplete set of measurements may be a member of only one of two classes, class E (earthquake) or class N (nuclear explosion). It will also be assumed that the cost of deciding that the set belongs to E when indeed it is a member of N (the cost of false dismissal) is c_1, and that the cost of deciding in favor of N when actually the set belongs to E is c_2 (the cost of false alarm).

If all the N-measurements on an input were available, and c_1 were equal to c_2, then a reasonable way of making classification decisions would be to decide using the following equation.

$$(x_1, x_2, --- x_N) \in E \text{ if } \frac{p(E|x_1, x_2, ---, x_N)}{p(N|x_1, x_2, ---, x_N)} > 1 \ . \qquad (35)$$

This decision states that if, given $x_1, x_2, \ldots x_N$, E is more likely than N , then one should decide that the N-dimensional vector $\vec{x}$ belongs to E. The opposite decision is made if the inequality does not hold. By applying Bayes' Rule, Equation 35 may be written as shown in Equations 36 and 37, where P_E and P_N are the a priori probabilities of occurrence of class E and N, respectively.

$$\vec{x} \in E \quad \text{if} \quad \frac{p(E|x_1, x_2, ---, x_N)}{p(N|x_1, x_2, ---, x_N)} = \frac{p_E(x_1, x_2, ---, x_N)P(E)}{p_N(x_1, x_2, ---, x_N)P(N)} > 1 \ , \qquad (36)$$

$$\frac{p_E(x_1, x_2, ---, x_N)}{p_N(x_1, x_2, ---, x_N)} = \ell(\vec{x}) > \frac{P(N)}{P(E)} = T = \text{const.} \qquad (37)$$

The function of $\ell(\vec{x})$ is the N-dimensional likelihood ratio. Several other likelihood ratios will be introduced later and will be distinguished from each other by subscripts which will indicate the decision rule with which the likelihood ratios are associated.

Suppose now that of the N-measurable parameters on which decisions should be based, only N-H are available. In the following, four of several possible decision rules are discussed for deciding, from available measurements which of the two classes E or N , is most likely.

Decision Rule One (Decisions Based on Marginal Densities)

This rule states that, given the measurements $(x_1, x_2, \ldots, x_{N-H})$, decide that these measurements belong to E if E is more likely than N as follows,

$$(x_1, x_2, ---, x_{N-H}) \in E \text{ if } \frac{p(E|x_1, x_2, ---, x_{N-H})}{p(N|x_1, x_2, ---, x_{N-H})} > 1 \ . \qquad (38)$$

by employing Bayes' rules, this equation can be rewritten as,

$$\frac{p(E|x_1, x_2, ---, x_{N-H})}{p(N|x_1, x_2, ---, x_{N-H})} = \frac{P(E)}{P(N)} \frac{p_E(x_1, x_2, ---, x_{N-H})}{p_N(x_1, x_2, ---, x_{N-H})} > 1 \ , \qquad (39)$$

or

$$(x_1, x_2, ---, x_{N-H}) \in E \text{ if } \frac{p_E(x_1, x_2, ---, x_{N-H})}{p_N(x_1, x_2, ---, x_{N-H})} > \frac{P(N)}{P(E)} = T \ . \qquad (40)$$

If we let

$$\frac{p_E(x_1, x_2, ---, x_{N-H})}{p_N(x_1, x_2, ---, x_{N-H})} = \ell_1 (x_1, x_2, ---, x_{N-H})$$

be the N-H dimensional likelihood ratio, abbreviated as $\ell_1(\vec{x})$ to simplify the notation, then decision rule 1 states that we should compare $\ell_1(\vec{x})$ with a threshold to determine if $\vec{x}$ should be classified in E or N . This, in effect, means that if $x_1, x_2, \ldots, x_{N-H}$ are the only measurements made, these measurements alone should be the basis for decision.

<u>Decision Rule Two</u> (Decisions Using Most Probable Values of Missing Measurements x_{N-H+1} through x_N)

After the N-H measurements $(x_1, x_2, \ldots, x_{N-H})$ are made, the probability density of the missing H measurements $(x_{N-H+1}$ through $x_N)$ can be calculated and the most probable values of the H measurements chosen for use in the N-dimensional likelihood ratio $\ell(\vec{x})$ can be determined. The value of the ratio $\ell(\vec{x})$ when the most probable values of the H missing measurements are used is $\ell_2(\vec{x})$. The most probable values are those which maximize the probability density given in Equation 41.

$$p(x_{N-H+1}, x_{N-H+2} \ldots x_N | x_1, x_2, ---, x_{N-H}) . \qquad (41)$$

Thus,

$$p(\hat{x}_{N-H+1}, \hat{x}_N | x_1 --- x_{N-H}) \geq p(x_{N-H+1}, ---, x_N | x_1, \ldots, x_{N-H}) . \qquad (42)$$

Accordingly, the decision rule states that one should decide

$$(x_1, ---, x_{N-H}) \in E \quad \text{if}$$

$$\ell_2(\vec{x}) = \frac{p_E(x_1, ---, x_{N-H}, \hat{x}_{N-H+1}, \ldots, \hat{x}_N)}{p_N(x_1, ---, \hat{x}_{N-H+1}, \ldots, \hat{x}_N)} > \frac{P(N)}{P(E)} = T \quad . \qquad (43)$$

This rule predicts the most likely values of the missing measurements and uses them as if they had actually been measured.

Decision Rule Three (Decisions Using the Most Probable Value of the Likelihood Ratio)

When only N-H measurements $(x_1, x_2, \ldots, x_{N-H})$ are made, the likelihood ratio $\ell(\vec{x})$ is a function of the unmeasured random variables x_{N-H+1} through x_N. This is denoted by $\ell_3(\vec{x})$ and is defined in Equation 44.

$$\ell_3(\vec{x}) = \frac{p_E(x_1, \ldots, x_{N-H}, x_{N-H+1}, \ldots, x_N)}{p_N(x_1, \ldots, x_{N-H}, x_{N-H+1}, \ldots, x_N)} \quad . \tag{44}$$

Accordingly, there is a probability density $p(\ell_3(\vec{x}))$ associated with $\ell_3(\vec{x})$ so that the most probable value of the likelihood ratio, given the observed $x_1, \ldots, x_{N-H}$ measurements can be determined. Thus,

$$p\left(\hat{\ell}_3(\vec{x})\right) \geq p\left(\ell_3(\vec{x})\right) \quad . \tag{45}$$

The decision rule is therefore, decide

$$(x_1, ---, x_{N-H}) \in E \text{ if } \hat{\ell}_3(\vec{x}) > T \ . \tag{46}$$

Decision Rule Four (Decisions Based on the Average Value of the Likelihood Ratio)

In this decision rule, the likelihood ratio is again treated as a function of the missing H measurements. However, instead of using the most probable value of $\ell_3(\vec{x})$ as in rule number 3, the average value of this likelihood ratio is used as the basis for deciding between E and N. Thus,

$$(x_1, ---, x_{N-H}) \in E \text{ if } \overline{\ell_3(\vec{x})} > T \ , \tag{47}$$

where

$$\overline{\ell_3(\vec{x})} = \int_{-\infty}^{\infty} \ell_3(\vec{x}) p[\ell_3(\vec{x})] d\ell_3 \tag{48}$$

Further, if $\ell_3(\vec{x})$ is a monotonic function of x_{N-H+1} through x_N , Equation 48 may be written as

$$\ell_3(x) = \int \cdots \int_{-\infty}^{\infty} \int \ell_3 \, (x_{N-H+1}, \cdots x_N) \cdot$$

$$p(x_{N-H+1}, \ldots, x_N | x_1, \ldots, x_{N-H}) dx_{N-H+1}, \cdots, dx_N \qquad (49)$$

To compare the different decision rules, the probabilities of error are computed and that rule which yields the smallest error probability is sought. The two error rates, the probability of false alarm and the probability of false dismissal, P(FA) and P(FD) , are given in Equations 50 and 51, where Y is the region in N-dimensional space in which the decision rule in question decides that the set of measurements $(x_1, x_2, \ldots, x_{N-H})$ belongs to N , and Y' is the region in which the decision favors class E . If there are only two classes, Y' is the complement of Y .

$$P(FA) = \int \int_{Y} \cdots p_E(x_1, x_2, \cdots, x_N) dx_1, dx_2, \cdots, dx_N \qquad (50)$$

$$P(FD) = \int \int_{Y} \cdots \int p_N(x_1, x_2, \cdots, x_N) dx_1, \cdots, dx_N \qquad (51)$$

However, given the values $(x_1, x_2, \ldots, x_{N-H})$, the likelihood ratios $\ell_1(\vec{x})$ through $\ell_3(\vec{x})$ are all functions of N-H given measurements alone. Thus, no matter how complicated the likelihood ratios may be, they are, for a specified choice of $P_E(x_1, \ldots, x_N)$ and $P_N(x_1, \ldots, x_N)$, deterministic functions of $x_1, x_2, \ldots, x_{N-H}$. Thus, the integrals of Equations 50 and 51 may be written as shown in Equations 52 and 53, where the region, y denotes the region of N-H dimensional space in which the measurement values are assigned to class N by the rule in question. Similarly, y' is the complement of y .

$$P(FA) = \int\int_{y} --- \int p_E(x_1, x_2, ---, x_{N-H})dx_1, dx_2, ---, dx_{N-H} \tag{52}$$

$$P(FD) = \int\int_{y'} --- \int p_N(x_1, x_2, ---, x_{N-H})dx_1, dx_2, ---, dx_{N-H} \tag{53}$$

If the positive constants c_1 and c_2 are the costs of false alarm and false dismissal, $p_E(x_1, ---, x_{N-H})$ and $p_N(x_1, ---, x_{N-H})$ are the marginal densities of the random processes E and N , and y_s and y_g are the regions in the N-H dimensional subspace of measured values in which decision rules s and g , respectively decide that the observations should belong to N , then rule s is better than rule g if the inequality of Equation 54 holds in the specified direction. Each side of the inequality expresses the probability of error according to the corresponding decision rule.

$$P(E)\ c_1 \int\int_{y_s} \cdots \int p_E(x_1, \ldots, x_{N-H})dx_1, ---, dx_{N-H} + P(N)\ c_2 \cdot$$

$$\int\int_{y'_s} \cdots \int p_N(x_1, ---, x_{N-H})dx_1, ---, dx_{N-H}$$

$$< P(E)\ c_1 \int\int_{y_g} --- \int p_E(x_1, ---, x_{N-H})dx_1, \ldots, dx_{N-H} + P(N)\ c_2 \cdot$$

$$\int\int_{y'_g} \cdots \int p_N(x_1, \ldots, x_{N-H})dx_1, \ldots, dx_{N-H} \tag{54}$$

Furthermore, given the densities $p_N(x_1, ---, x_{N-H})$ and $p_E(x_1, ---, x_{N-H})$, it is seen that the decision rule which minimizes Q is best.

$$Q = P(E)\ c_1 \int\int_{y_s} \cdots \int p_E(x_1, \ldots, x_{N-H})dx_1, \ldots, dx_{N-H}$$

$$+ P(N)\ c_2 \int\int_{y'_s} \cdots \int p_N(x_1, \ldots, x_{N-H})dx_1, \ldots, dx_{N-H} \qquad (55)$$

However, since there are only two classes, Equation 56 allows simplification of Equation 55 to Equation 57.

$$\int\int_{y'_s} \cdots \int p_N(x_1, \ldots, x_{N-H})dx_1, \ldots, dx_{N-H} = 1$$

$$- \int\int_{y_s} \cdots \int p_N(x_1, \ldots, x_{N-H})dx_1, \ldots, dx_{N-H} \qquad (56)$$

$$Q = c_2\ P(N) + \int\int_{y_s} \cdots \int [c_1\ P(E)\ p_E(x_1, \ldots, x_{N-H})$$

$$- c_2\ P(N)\ p_N(x_1, \ldots, x_{N-H})]\ dx_1, \ldots, dx_{N-H} \qquad (57)$$

It is seen that Q is smallest of y_s is in the region in which the integral is always negative. In this region,

$$c_2\ P(N)\ p_N(x_1, \cdots, x_{N-H}) > c_1\ P(E)\ p_E(x_1, \cdots, x_{N-H}) \ . \qquad (58)$$

Further, for the case where $c_1 = c_2$, this reduces to the decision rule; decide N if

$$\frac{p_N(x_1, x_2, ---, x_{N-H})}{p_E(x_1, x_2, ---, x_{N-H})} > \frac{P(E)}{P(N)} = T . \qquad (59)$$

We recognize that this is just the decision made by the marginal densities of decision rule 1.

Thus, it is seen that the optimum classification decision based on N-H observed measurements of the set of N measurements consists of comparing the ratio of a posteriori probabilities of these observed measurements with a threshold and that no useful purpose is served by knowing to what higher dimensional process the N-H measurements belong.

REFERENCES

1. N. Abramson and D. Braverman, "Learning to Recognize Patterns in a Random Environment", IRE Transactions on Information Theory, Vol. IT-8, pp. 58-63, Sept. 1962.

2. G. Sebestyen, "Recognition of Membership in Classes", IRE Transactions on Information Theory, Vol. IT-6, pp. 44-50, January 1961.

3. G. Sebestyen, "Decision-Making Processes in Pattern Recognition," The Macmillan Company, New York, 1962.

4. G. Sebestyen, "Pattern Recognition by an Adaptive Process of Sample Set Construction", IRE Transactions on Information Theory, Vol. IT-8, No. 5, pp. S82-S91, Sept. 1962.

5. E. Glaser, "Signal Detection by Adaptive Filters", IRE Transactions on Information Theory, Vol. IT-7, pp. 87-98, April 1961.

6. G. H. Ball, The Application of Integral Geometry to Machine Recognition of Visual Patterns, WESCON, 1962.

7. W. H. Highleyman, "The Design and Analysis of Pattern Recognition Experiments," Bell System Technical Journal, March 1962, pp. 723-744.

8. A. Mood, Introduction to the Theory of Statistics, McGraw-Hill, New York, 1950.

9. P. M. Lewis, "The Characteristic Selection Problem in Recognition Systems," IEEE Transactions on Information Theory, Vol. IT-8, No. 2, Feb. 1962, pp. 171-178.

10. W. H. Highleyman, "Linear Decision Functions with Application to Pattern Recognition," Proc. IRE, pp. 1501-1514, June 1962.

11. D. Middleton and D. Van Meter, "On Optimum Multiple-Alternative Detection of Signals in Noise," IRE Transactions on Information Theory, Vol. IT-1, No. 2, Sept. 1955.

12. W. W. Peterson, T. G. Birdsall, and W. C. Fox, "The Theory of Signal Detectability", IRE Transactions on Information Theory, Vol. IT-1, No. 3, 1955.

13. D. W. Y. Sommerville, "An Introduction to the Geometry of N-Dimensions", Ch. VIII, Dover Publications, Inc., New York, N.Y., 1958.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

The MITRE Corporation
Bedford, Massachusetts

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

A SEISMIC CLASSIFICATION MODEL

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(First name, middle initial, last name)*

CLARK, Jack W.

6. REPORT DATE

September 1967

7a. TOTAL NO. OF PAGES

55

7b. NO. OF REFS

13

8a. CONTRACT OR GRANT NO.

AF 19(628)-5165

b. PROJECT NO.

6050

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

ESD-TR-67-75

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

MTR 305

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY Directorate of Planning and Technology Development Engineering Division, Electronic Systems Division, L. G. Hanscom Field, Bedford, Mass.

13. ABSTRACT

This report is intended as an introduction to one possible approach to the seismic classification problem. It develops a very general classification model using automatic non-parametric learning based on limited data of known classification. The model accepts discriminants extracted from the seismogram and yields the probability that the input was due to an earthquake or an explosion. Thus, the discriminants are assumed to be available as inputs. Pattern recognition as used here is defined, the classification procedure is outlined, the adaptive estimation of joint probability-densities from a finite number of multi-dimensional vectors of known classification (the learning model) is discussed, a simplified flow diagram of the learning model is presented, and the selection of necessary control parameters is investigated.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Seismic Classification						
Seismograms						
Earthquakes and Explosions						
Probability Densities						